

А.М. Лях  
**МИНИМАЛЬНАЯ СТРУКТУРА БАЗЫ ДЛЯ ХРАНЕНИЯ  
ДАННЫХ О БИОЛОГИЧЕСКОМ РАЗНООБРАЗИИ  
ОРГАНИЗМОВ**

*Лях А.М. Минимальная структура базы для хранения данных о биологическом разнообразии организмов.*

**Аннотация.** К настоящему времени накоплено огромное количество данных о разнообразии организмов. Сохранить и использовать эти данные для решения научных задач помогают базы данных. В литературе описано несколько десятков баз, предназначенных для хранения данных о биоразнообразии. Каждая имеет оригинальную структуру, которая плохо согласуется со структурами других баз, что, в свою очередь, затрудняет обмен данными и формирование массивов больших данных о биоразнообразии.

Причиной сложившейся ситуации является отсутствие формальных определений универсальных компонент, из которых можно построить базу с любыми данными о разнообразии организмов. Анализ литературы и исследования авторов показывают, что универсальные компоненты есть в характеристиках любых организмов. Например, таксономическое название организма и место его поимки. Таких компонент шесть, и они отвечают на один из шести вопросов: *что, где, когда, кто, откуда и куда*. Первые три компонента – *что, где, когда* – являются фундаментальными. Они составляют минимальную основу, которая описывает экземпляр таксона в пространственно-временных координатах. Каждой компоненте соответствует отдельная таблица базы данных. Эти таблицы связаны с таблицей данных об организме (особи) и не связаны между собой. Атрибуты связей между особью и таблицами компонент хранятся в промежуточных таблицах.

Процесс создания любой базы о разнообразии живых существ начинается с определения таблицы экземпляров организмов. Ее необходимо использовать, даже если нет явных данных об организмах. Тогда следует ввести виртуальные организмы и связать с ними остальные компоненты при помощи промежуточных таблиц. Последние стыкуются с прочими данными. Минимальные структуры всех таблиц, связи между ними и примеры построения баз данных описаны в настоящей работе.

**Ключевые слова:** компоненты данных, данные об организме, географическая точка, таксономическое название, библиографическая запись, биологическая коллекция, коллекционный экземпляр, метаданные.

**1. Введение.** Под биологическим разнообразием (биоразнообразием) понимают совокупность всех живых существ и групп организмов, населяющих определенную территорию (акваторию) или биосферу в целом. Выделяют несколько уровней биоразнообразия: генетический, индивидуальный, популяционный, таксономический, биоразнообразие сообществ и тому подобное [1, 2]. Но единственным фундаментальным уровнем биоразнообразия является разнообразие живых организмов, так как только естественные или искусственно выделенные группы организмов создают разнообразие иных уровней жизни.

Исследование биоразнообразия начинается с инвентаризации накопленных данных, которая приводит к обнаружению новых фактов

о живых существах. Совокупность добытой информации необходимо где-то хранить, желательно в структурированном виде, удобном для извлечения данных и удовлетворяющем разнообразным критериям. Такие выборки помогают решить научные задачи инвентаризации, анализа, моделирования и прогноза биоразнообразия, и следовательно, способствуют выработке рекомендаций для его сохранения. Лучше всего задачу структурированного хранения и многокритериального выбора данных помогают решить базы данных [1, 3, 4].

Существует несколько крупных хранителей данных о биоразнообразии – это обладающие богатейшими биологическими коллекциями организации, такие как МГУ (Москва) [5], Зоологический институт РАН (Санкт-Петербург) [6], Музей естественной истории (Лондон); либо организации, которые длительное время занимаются экологическим мониторингом [7]. Некоторые из них выступают с идеей объединить всю информацию о биоразнообразии в единое общедоступное информационное пространство и предлагают для этого собственные площадки [5, 6, 8]. Например, Глобальная база данных по объектам биоразнообразия (англ., Global Biodiversity Information Faculty; GBIF) предлагает исследователям разместить данные о находках организмов в глобальной базе данных [9], а взамен предоставляет унифицированный перечень полей для описания собранного материала [10] и обеспечивает доступ к данным остальных участников. Однако до сих пор существует большое число оригинальных баз данных, которые принадлежат отдельным организациям и не связаны в единую систему [11-15]. Разработчики этих продуктов чаще всего стараются решить локальную задачу и не планируют в дальнейшем включать в базу новые модули или объединять ее с другими базами. В результате накопленные данные остаются неизвестными научному сообществу, и существует опасность, что они со временем будут безвозвратно утеряны [16, 17].

Таким образом, данные о биоразнообразии востребованы, но для их детального анализа, моделирования и прогноза необходимы большие массивы. Такие массивы могут предоставить организации, которые поддерживают глобальные базы данных. Однако информационная инфраструктура, обеспечивающая работу с данными, не позволяет эффективно решить частные исследовательские задачи. Во-первых, предлагаемые форматы содержат массу ненужных для частной задачи сущностей; во-вторых, стандартизацией предлагаемых форматов занимаются большие коллективы, из-за чего внесение любых правок затянато, так как проходит длительную бюрократическую процедуру согласований; в-третьих, глобальные хранилища предполагают однократную загрузку данных, а не их динамическую модификацию; в-четвертых, переданные данные

становятся общедоступными, что нежелательно для владельца первичных, еще необработанных и неопубликованных данных. По этим и иным причинам небольшие коллективы продолжают придумывать новые структуры баз [11-15], которые оптимально подходят для текущих исследований, но плохо стыкуются друг с другом. Чтобы хоть как-то решить проблему существования огромного числа локальных БД, некоторые исследователи предлагают создать аннотированный каталог всех баз данных по биоразнообразию [17]. Но в целом задача формирования большого общедоступного массива накопленных данных об организмах остается нерешенной [17, 18].

Предположительно, основной причиной сложившейся ситуации является отсутствие формального определения универсальных компонент, из которых можно построить базу с любыми данными о разнообразии организмов. Опыт авторов по разработке многочисленных баз биологических данных и анализ существующих работ по теме показали, что универсальные компоненты можно обнаружить во всех данных, связанных с исследованием организмов. Например, в библиотеках изображений особей, в таксономических списках, в базах морфологических и морфометрических признаков и молекулярно-генетических последовательностей, в виртуальных биологических коллекциях, в базах с результатами биологического мониторинга, в библиографии с описаниями групп таксонов. Все эти коллекции содержат дополнительную информацию об организме. Так как вариантов для представления такой информации немного, то эти описания включают общие части, например название, место и время находки. Именно эти части являются элементарными кирпичиками, которые составляют фундамент любой базы данных о биоразнообразии.

Цель работы – описать структуру и связи между элементарными компонентами, из которых состоит база данных о биоразнообразии, и рассмотреть практические приемы использования этих компонент.

**2. Данные о биоразнообразии.** Данные о биологическом разнообразии, в первую очередь, основаны на данных об отдельном организме (особи, экземпляре, индивидууме), так как не может быть биоразнообразия без организма. При изучении разнообразия видов, генетического разнообразия, разнообразия морфологических форм и экологических группировок всегда используются первичные данные, полученные в процессе исследования конкретных особей. Поэтому характеристики экземпляров организмов являются ядром любой базы данных (БД) о биоразнообразии.

Практический опыт разработки информационных систем по биоразнообразию и анализ большого числа данных показывают, что в

характеристиках любых организмов есть общая часть. Ее содержимое делится на шесть компонент. Удобно считать, что состав компонент содержит ответ на один из вопросов: *что, где, когда, кто, откуда и куда*. *Что* – определяет название организма; *где* – место находки; *когда* – время находки; *кто* – перечисляет тех, кто имеет отношение к обнаружению и исследованию организма; *откуда* – ссылается на научные источники, откуда были взяты или где были опубликованы данные об организме; *куда* – показывает, в какую коллекцию помещен организм. Данные вопросы являются мнемоническими правилами для запоминания элементарной структуры базы данных.

Первые три компонента – *что, где, когда* – являются фундаментальными. Они составляют минимальную основу, которая описывает экземпляр таксона в пространственно-временных координатах [2]. Остальные компоненты описывают детали находки и первичного анализа особи, уточняют источники дополнительных данных и помогают отыскать экземпляр организма в материальной или виртуальной коллекциях.

**3. Минимальная структура базы данных о биоразнообразии.** База данных о биоразнообразии строится из элементарных компонент, каждой из которых соответствует отдельная таблица. Эти таблицы связаны с таблицей данных об особи и не связаны между собой, что подчеркивает нацеленность базы на хранение данных об организме (рис. 1).



Рис. 1. Компоненты минимальной структуры базы данных о биоразнообразии; промежуточные таблицы отмечены кружками

Компоненты связаны с особью отношениями один-ко-многим или многие-ко-многим. Для описания отношений между особью и

компонентами используются промежуточные таблицы. Эти таблицы также функционируют как стыковочные станции, к которым присоединяются таблицы с прочей информацией. Таким способом можно расширять минимальную структуру базы данных.

Структуры таблиц, связи между ними и примеры использования описаны ниже.

**4. Универсальные компоненты данных об организме. Таблица с данными об особи.** Характеристики живых организмов хранятся в *таблице экземпляров* (на всех рисунках обозначена как *особь*). Любой организм можно описать огромным количеством параметров. Чем более точные методы исследования мы используем, тем больше данных об организме получаем. По этой причине проблема анализа больших данных также актуальна и для данных об организме [19].

Одной из теоретических оценок возможного числа данных об организме может служить число комбинаций, которые возможно составить из всех элементарных частиц, слагающих организм. Эти данные не учитывают другие характеристики живого существа, например поведение, предпочитаемые экологические условия, взаимоотношения с другими особями. Поэтому теоретическую оценку количества разнообразных данных об организме можно смело увеличить еще на несколько порядков.

К счастью, в реальных условиях количество признаков, которые используются для характеристики организма, невелико [15]. Предлагаемая таблица с минимальной структурой описания особи содержит первичный ключ (*specm\_id*) и три группы полей для фундаментальных компонент (рис. 2).

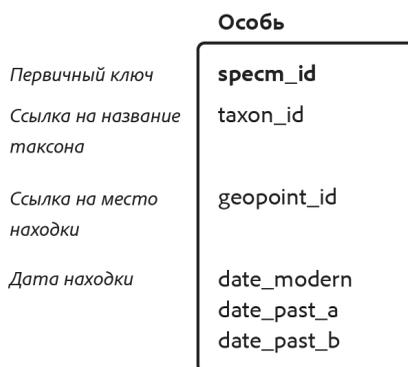


Рис. 2. Минимальная структура таблицы для описания экземпляров организмов (особей) содержит указатели на фундаментальные компоненты: название таксона, место и дату находки организмов

Все прочие данные об организме факультативны. Они хранятся или в дополнительных полях, или в других таблицах, связанных с особью, и в настоящей работе не рассматриваются.

**4.1. Компонент *что* – таблица с научными названиями организмов.** Все известные науке организмы отнесены к некоторым таксонам и названы таксономическими именами. Таксон – это множество организмов, объединенных по некоторым признакам. Таксоны последовательно вложены друг в друга и в совокупности образуют иерархически упорядоченное дерево. Листьями дерева таксонов являются естественные виды организмов. Не теряя общности, можно считать, что дерево является системой иерархически связанных названий таксонов, потому что любое таксономическое название – это ключ для доступа к данным об организмах.

Таблица с названиями организмов хранит дерево таксономических имен. Для удобства структура таблицы разделена на три смысловые группы полей (модулей): иерархическую, номенклатурную и атрибутивную [20] (рис. 3).

	Что Название таксона	Где Географическое место
<i>Первичный ключ</i>	taxon_id	place_id
<i>Иерархическое положение</i>	parent_id tree_level tree_path	parent_id tree_level tree_path
<i>Номенклатурные данные</i>	taxon_name taxon_author taxon_year	place_name
<i>Атрибуты</i>	status	status

Рис. 3. Структуры таблиц для названий таксонов и географических названий

*Иерархическая* группа полей описывает положение названия в таксономическом дереве. Для этого используется структура данных *материализованный путь* [20, 21]. Эта структура компактна и наглядна – для ее описания достаточно двух полей: путь к узлу и относительный уровень узла, который одновременно кодирует ранг таксона. Такая структура дает возможность хранить неполные ветви, в которых пропущены названия промежуточных уровней (рангов). Она также позволяет выбирать названия любого ранга, которые относятся к лю-

бому вышестоящему узлу, например названия видов из заданного класса или отряда.

*Номенклатурная* группа полей служит для хранения полного таксономического названия организма, удовлетворяющего требованиям таксономической номенклатуры. Она включает поля для названия таксона, авторов таксономического названия и года обнародования названия.

*Атрибутивная* группа полей включает одно числовое поле. Оно показывает статус таксономического имени и указывает на действующее (валидное) название, если данное название не валидно. Валидным является название, которое действует сейчас и удовлетворяет правилам одного из кодексов таксономической номенклатуры. Нулевое значение поля обозначает валидное название. Отрицательное – не валидное название, которое составлено или опубликовано с нарушением правил. Положительное значение сообщает, что данное название – синоним другого названия и это значение является ссылкой на идентификатор записи с действующим названием таксона.

Указанных трех групп полей достаточно для полноценного представления и работы с деревом таксономических имен [20].

**4.2. Компонент где – таблица с географическими названиями мест находки организмов.** Для предварительного указания места обнаружения организма достаточно географического названия территории. Оно понятно любому человеку лучше географических координат. Для хранения географических названий используется отдельная таблица, которая одновременно служит своеобразным справочником.

Географические названия, так же как и названия таксонов, образуют древовидную иерархию. Ее корнем является вся Земля. Она подразделяется на Мировой океан и материки, вплоть до конкретной географической точки. Так как в географии нет устоявшейся иерархии уровней, то выстроить однозначную ветвь (цепочку) географических названий достаточно сложно. С другой стороны, это не столь необходимо. В цепочке географических имен нам всегда интересна конечная точка и несколько названий верхних уровней. Например, из цепочки «бухта Провато – акватория Крыма – Черное море – (Средиземное море – Атлантический океан – Мировой океан)» можно понять, что мы говорим о бухте в Черном море на побережье Крыма. Промежуточные уровни в скобках понадобятся, когда мы захотим выбрать все морские (океанические) точки или все точки из Средиземного моря или Атлантического океана, так как Черное море относится к бассейну этих акваторий. Но и без них указание места будет понятным.

Для хранения иерархии географических имен подходит та же структура модульной таблицы БД, которая используется для названий таксонов [20]. В этой таблице иерархическая группа полей остается прежней, номенклатурная группа включает одно поле для названия местности, а поле атрибутивной группы указывает на синонимы географических названий так же, как и в случае синонимов таксономических имен (рис. 3).

**4.3. Компонент *когда* – дата находки организма.** Любой организм всегда пойман в одно конкретное время. Поэтому для записи времени не требуется дополнительная таблица – оно хранится непосредственно в поле таблицы экземпляров.

Однако ученые исследуют не только современное биоразнообразие, но и биоразнообразие прошлых эпох по палеонтологическим находкам [11, 22]. Датировка любых ископаемых образцов всегда примерна и обычно описывается интервалом. Поэтому для хранения диапазона ископаемых дат используется еще два поля. Так как палеонтологические даты исчисляются сотнями тысяч и миллионами лет, то они записываются в экспоненциальной форме.

Таким образом, для записи дат используется три поля: одно для точной современной даты (*date\_modern*), два – для примерного датирования ископаемых образцов (*date\_past\_a*, *date\_past\_b*). Даты хранятся в таблице с описанием организмов (особей) (рис. 2).

**4.4. Компонент *кто* – таблица с именами исследователей.** При изучении биоразнообразия имя исследователя предоставляет определенную информацию об особенностях сбора, обработки и таксономической идентификации и анализе собранных данных. В каждой области научной деятельности есть признанные специалисты, есть начинающие исследователи и есть люди, запятнавшие свою репутацию. Поэтому, например, если записано, что виды определил исследователь А, то у специалистов может возникнуть сомнение в правильности этой идентификации, потому что большая часть его определений неверная. Или если указано, что организмы для анализа подготовил исследователь Б, то сообществу исследователей станет понятно, что его материал никоим образом нельзя сопоставлять с материалом исследователя В, так как они применяют несопоставимые методики. Таким образом, имя исследователя в данном случае является косвенным показателем качества данных.

Имена всех исследователей хранятся в соответствующей таблице имен. Имя записывается в трех полях для имени (*name*), фамилии (*family*) и отчества или прочих частей имени (*patronym*). Однако такой записи недостаточно для того, чтобы различить полных

тезок. В научном сообществе полных тезок можно отличить, например, по городу или названию организации, в которой работал человек. Если и этого недостаточно, можно использовать какую-то иную характеристику каждого человека, понятную остальным. Например, примерный возраст (старший, младший), специализацию (ихтиолог, микробиолог, малаколог, ботаник), должность. Для этих характеристик предназначено дополнительное поле (*nick*) (рис. 4).

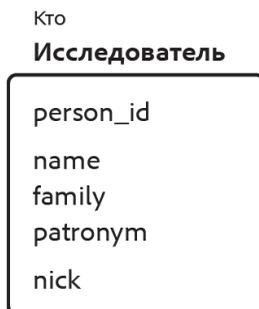


Рис. 4. Минимальная структура таблицы для хранения имен исследователей

Все остальные поля таблицы являются необязательными, а состав полей определяется конкретным назначением базы данных. Например, в таблицу можно включить поля с контактными данными человека. Однако, если база содержит исторические данные и кого-то из исследователей нет в живых, то их контактные данные будут представлять только историческую ценность.

**4.5. Компонент *откуда* – таблица с публикациями.** Публикации являются богатейшим источником неструктурированной информации. Крайне важно эту информацию разделить на отдельные факты и связать с организмами, которые упоминаются в публикациях. Но прежде необходимо сохранить данные о публикациях.

Библиографическая таблица хранит библиографические ссылки на публикации, а не сами публикации, потому что многие из них охраняются авторским правом. Краткая библиографическая запись состоит из фамилий и инициалов авторов (*authors*), заголовка (*title*), года опубликования (*year*), прочих выходных данных (*imprint*) и DOI (*doi*). Для описания сетевой публикации без DOI используется URL (*url*) и дата последнего обращения к документу (рис. 5). Это минимальный набор полей библиографической записи, которые можно дополнять прочими данными.

Откуда

**Публикация**

publ_id
authors
title
year
imprint
doi
url
request_date

Рис. 5. Минимальная структура таблицы для хранения библиографических записей

#### 4.6. Компонент *куда* – таблица с описанием коллекций.

Биологические коллекции служат для длительного сохранения биологических объектов [23, 24], к которым в том числе относятся организмы или части организмов. Биологическая коллекция является одним из основных источников данных для исследования биоразнообразия.

Полноценное исследование биоразнообразия подразумевает депонирование найденных организмов в биологические коллекции. Это позволяет повторно использовать любую особь для уточнения морфологических, генетических, морфометрических и прочих признаков, на которые исследователь при первичном анализе организма не обратил внимание. Таким образом, биологические коллекции предотвращают утрату ценной научной информации.

Минимальная структура таблицы с описанием биологической коллекции (рис. 6) включает только уникальный (буквенно-цифровой) идентификатор (*coll\_abbr*). Подобные идентификаторы (аббревиатуры) присвоены некоторым коллекциям (см. например [25]). Они помогают кратко сослаться на коллекцию, а также используются для формирования уникальных номеров коллекционных экземпляров [26]. Любая другая информация о названии коллекции, организации, которая содержит коллекцию, ответственных за коллекционные материалы, веб-адресе виртуального коллекционного ресурса и прочие сведения, является факультативной. Эти сведения дополняют минимальную структуру таблицы и в настоящей работе не рассматриваются.

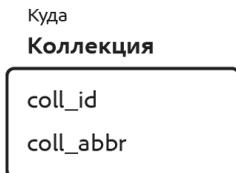


Рис. 6. Минимальная структура таблицы для описания биологической коллекции

**5. Отношения между компонентами и особью.** Особь связана с компонентами при помощи отношений. Свойства этих отношений описаны в дополнительных таблицах. Рассмотрим каждую из них подробнее.

**5.1. Отношение что: особь – название таксона.** Любому организму соответствует конкретное название таксона. Чаще всего это название вида, так как любой организм всегда стремится идентифицировать до вида. Однако даже если мы не знаем название вида, то вполне можем отнести особь к таксону более высокого ранга: роду, семейству, классу или даже царству, и впоследствии уточнить результаты определения. Например, любой образованный рыбак, поймавший барабульку у берегов Ялты, сразу же скажет ее видовое название *Mullus barbatus*, тогда как менее образованный только предположит, что это представитель отряда *Perciformes*. Таким образом, оба рыбака установили отношение между пойманным экземпляром рыбы и таксономическим названием вида или отряда.

Всегда существует однозначная связь особи с конкретным таксономическим названием. По этому названию с помощью дерева названий таксонов можно получить имена таксонов вышестоящих рангов. Поэтому в таблице экземпляров поле для названия организма содержит только ссылку на запись (идентификатор записи) с названием таксона из соответствующей таблицы (рис. 7).

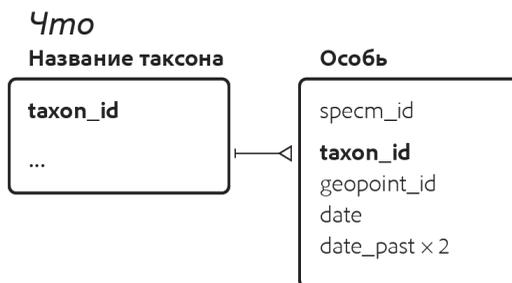


Рис. 7. Отношение между таксономическим названием и особью

**5.2. Отношение *когда*: особь – дата.** Это однозначное отношение и промежуточной таблицы для его хранения не требуется, так как дата записана в таблице особей (рис. 2).

Другим вариантом представления является выделение отдельной таблицы для палеонтологических дат. Такие даты относят к определенным эпохам: эоцен, миоцен, плейстоцен и так далее. Таблица будет хранить название эпохи и диапазон дат (тысячи или миллионы лет), определяющих эту эпоху.

**5.3. Отношение *где*: особь – точка на карте – географическое место.** Для точного описания места поимки организма недостаточно географического названия местности. Нужны дополнительные данные. К ним относятся, в первую очередь, географические координаты, а также ориентир и высота или глубина места поимки. Так как эти данные нельзя поместить в таблицу с названиями географических мест, для них используется промежуточная таблица. Она связывает особь с точкой на карте, а точку – с географическим местом. Эти связи многозначные, потому что в одной точке может быть найдено много особей и в одном районе может быть много точек (рис. 8).

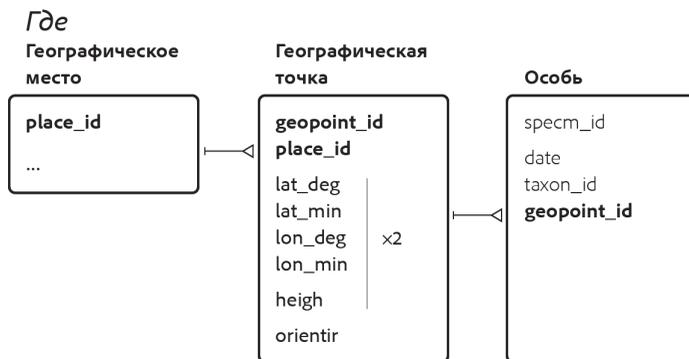


Рис. 8. Отношение между географическим местом, географической точкой и особью; ×2 означает, что указанные поля удвоены

Таблица для описания точки включает поля для географических координат, высоты или глубины и словесного описания ориентиров. Для координат предназначено восемь полей: две группы по четыре поля. В каждой группе два поля хранят градусы широты и долготы (*lat\_deg*, *lon\_deg*), два – десятичную запись минут (*lat\_min*, *lon\_min*), где отрицательные значения соответствуют южной широте и западной долготе. Две координаты позволяют записать координаты горизонтальных траловых сборов. Если обе координаты совпадают или вторые пустые – это единственная точка.

Высоты или глубины также хранятся в двух полях (*height*), где глубине соответствует отрицательное значение. Когда значения в обоих полях совпадают, они описывают одну точку; когда отличаются – вертикальный слой глубин или высот. Такие данные получаются, например, при сборе планктонных проб. Если заданы две координаты и первой координате соответствует одна глубина, а второй – другая, то такие данные опишут наклонный путь.

Ориентир записывается в отдельном поле (*orientir*) в свободной форме. Например, «проба собрана у родника в пяти метрах на северо-восток от входа в пещеру». Ориентиры помогают найти точку сборов, когда точные координаты неизвестны или их невозможно определить.

**5.4. Отношение *кто: особь* – роль исследователя – исследователь.** Каждый исследователь выполняет определенную роль при сборе особей. Один – отлавливает организмы, другой – определяет вид, третий – выполняет молекулярно-генетический анализ, четвертый – готовит постоянные препараты для коллекции. Очевидно, при сборе разных организмов меняются как роли, так и сами исследователи. Также один исследователь может сразу выполнять несколько функций. Например, исследователь С, в одном случае ловил лягушек, в другом – ловил и определял, в третьем – подготавливал для длительного хранения в коллекции, в четвертом – вместо него все делал исследователь З. Роли исследователей, которые принимали участие в сборе экземпляров, записываются в таблице ролей (рис. 9). В этой же таблице можно предусмотреть поля для уточнения методики сбора или анализа, для важных комментариев и прочей информации.

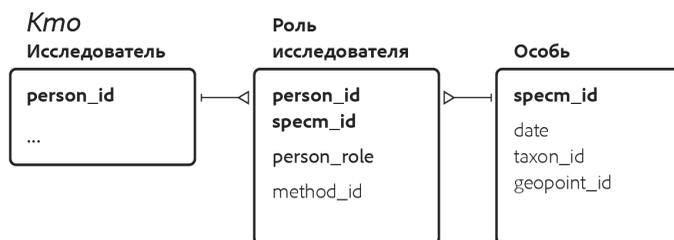


Рис. 9. Отношения между исследователем, ролью исследователя и особью

У каждого исследователя может быть несколько ролей и одной особью могут заниматься несколько исследователей, поэтому отношения между этими таблицами многие-ко-многим. Так как роли будут постоянно повторяться, то можно дополнительно хранить справочник ролей в отдельной таблице (на рисунке не показана). Таковую же таблицу можно предусмотреть для словаря с описанием методик.

**5.5. Отношение *откуда*: публикация – факт – особь.** Из публикаций мы черпаем факты об организмах и в публикациях мы описываем новые факты. Фактом является предложение, абзац или параграф с текстом, изображение, табличные данные. Вот, например, факт о строении представителей какого-то вида морских свободноживущих нематод: «губернакулом окружает дистальный конец спикул с дорсально направленным апофизом 30 мкм длиной».

Факты описаны в таблице фактов (рис. 10). Кроме самих фактов (*fact*) она хранит ссылки на них (*fact\_ref*). Это может быть номер рисунка или таблицы, страница статьи или отсылка к конкретному параграфу. В целом ссылка уточняет, откуда и какие именно данные были извлечены из публикации.

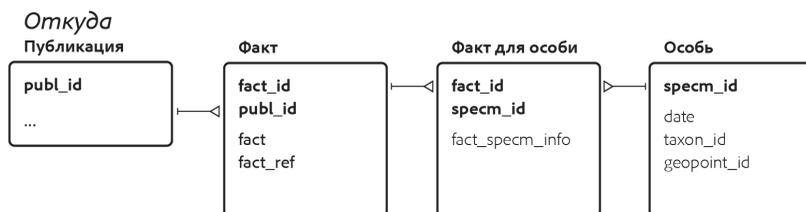


Рис. 10. Отношение между публикацией, фактом и особью

Так как любой факт может относиться к нескольким особям и каждая особь характеризуется множеством фактов из разных публикаций, то для связи особи с фактом предусмотрена еще одна промежуточная таблица (рис. 10). В ней также может быть записана информация (*fact\_specm\_info*), которая уточняет факт для данного организма.

**5.6. Отношение *куда*: особь – коллекционный экземпляр – коллекция.** Ценность любой биологической коллекции составляют экземпляры живых организмов [24]. Но таблица с описанием коллекций (рис. 6) не хранит данные о коллекционных экземплярах. Для этого предназначена вспомогательная таблица.

Минимальная структура таблицы с коллекционными экземплярами состоит из уникального коллекционного номера единицы хранения (*specm\_numb*), идентификатора коллекции (*coll\_id*) и особи (*specm\_id*), типового статуса сохраненного экземпляра (*type\_status*), сведений о цельности коллекционного образца (*specm\_integrity*), о стадии развития (*ontogeny\_stage*) и о состоянии хранимого объекта (*specm\_condition*) (рис. 11) [14].

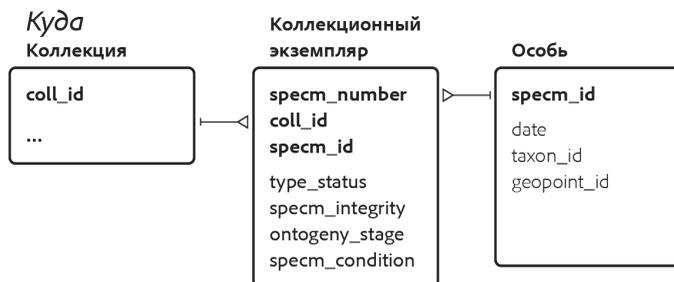


Рис. 11. Отношения между коллекцией, коллекционным экземпляром и особью

Для однозначной идентификации любого коллекционного экземпляра используют коллекционные номера [26, 27]. Коллекционный номер – это уникальная символьная метка, привязанная к коллекционному экземпляру. Она нужна для однозначной ссылки на конкретный организм при проведении научных исследований. Номер физически наносится на коллекционный экземпляр или препарат с экземпляром. Его записывают на каталожной карточке с информацией о единице хранения и используют в качестве идентификатора записи коллекционной базы данных. Таким образом, коллекционный номер служит для однозначной идентификации материальных коллекционных носителей и виртуальных информационных образов этих носителей. В таблице с коллекционными экземплярами коллекционный номер одновременно служит первичным ключом.

С одним организмом может быть связано несколько коллекционных экземпляров с разными номерами. Такая ситуация возникает, когда в коллекции хранятся части некогда целого организма и каждая часть считается отдельной единицей хранения. Например, коллекция костей стеллеровой коровы содержит множество костей некогда целого скелета, каждая из которых имеет уникальный номер и является отдельной единицей хранения [28]. Поле «цельность организма» (*specm\_integrity*) как раз сообщает, часть или целый организм скрывает данный номер, а идентификатор особи (*specm\_id*) позволяет выбрать все части одного организма.

С коллекционным номером соотнесен типовой статус экземпляра (*type\_status*): голотип, паратип, неотип, нетиповой (ваучерный) экземпляр. Отдельное поле (*ontogeny\_stage*) описывает стадию развития (онтогенеза) организма: взрослая особь, личинка, яйцо и тому подобное. Также предусмотрено поле (*specm\_condition*) для данных о состоянии сохраненного объекта: отличное, хорошее, плохое.

Прочие сведения о возрасте и поле особи, типе сосуществования с другими организмами, экологических предпочтениях, не являются общими для всех живых существ и относятся к факультативным данным [14], поэтому в работе они не рассматриваются.

**6. Метаданные о данных о разнообразии организмов.** Метаданные – это данные о данных, краткий перечень того, что содержит база данных. Метаданные помогают выбрать из огромного разнообразия баз только те, которые содержат необходимые для исследования сведения [29]. Метаданные используют как люди, так и машины. Последние – для генерации метабаз с данными о базах данных, например, метабаза о данных о биоразнообразии [17]. Несомненно, метабаза данных удобно использовать для многокритериального поиска информации.

Предложенные компоненты минимальной базы данных о биоразнообразии (рис. 1) подходят для автоматической генерации метаданных. Они включают статичные фундаментальные компоненты с названием высших таксонов и крупных географических провинций, где собраны данные [30], краткую информацию о коллекциях, куда депонированы экземпляры организмов, и библиографические сведения о публикациях, где опубликованы данные об этих экземплярах. Помимо этого, метаданные могут содержать динамические данные о числе особей, которые собраны в указанном районе, в указанное время, депонированы в указанную коллекцию и относятся к указанному таксону. Такие метаданные генерируются в ответ на запрос и достаточно полно описывают содержимое базы данных.

**7. Использование компонент для построения баз данных.** Процесс создания любой базы о разнообразии живых существ начинается с определения таблицы экземпляров организмов. Ее необходимо использовать, даже если нет явных данных об организмах. Тогда следует ввести виртуальные организмы и связать с ними остальные компоненты при помощи промежуточных таблиц. Такую минимальную структуру базы данных можно расширить путем добавления новых полей к существующим таблицам или новых таблиц.

Ниже даны примеры, которые показывают, как из компонентов построить две базы с биологическими данными. Первая база помогает хранить список таксонов. Вторая предназначена для хранения описаний организмов, собранных во время морских экспедиций и депонированных в биологическую коллекцию.

**7.1. Список видов, населяющих территорию.** Флористические и фаунистические списки перечисляют виды организмов населяющих определенную территорию. В списке обычно присутствует два компонента: названия видов (компонент *что*) и названия местно-

сти (компонент *где*), и нет явных данных про экземпляры организмов и время находки (компонент *когда*).

Простейшим очевидным описанием списка таксонов является база, которая состоит из двух элементарных компонент: таблицы с названием таксонов связанной с таблицей с названием территорий (рис. 12а). Однако, эта структура базы является не оптимальной, так как ее центральный элемент – таблица с названиями таксонов. По этой причине такую базу данных можно дополнить только модулями, которые содержат информацию о таксонах – то есть о группе организмов, а не об отдельных организмах. Однако первичными данными для таксономических списков всегда является информация об экземплярах организмов, а не о таксонах. Следовательно, сохранить эту первичную информацию в данной базе не получится.

Правильное решение заключается в том, чтобы ввести в базу таблицу с *виртуальными экземплярами* и связать с ними все названия таксонов и территорий (рис. 12б). Виртуальные экземпляры – это особи, которые когда-то были найдены в данной местности, но подробности этих находок (и, скорее всего, сами особи) не сохранились. Известны только их таксономические названия. С помощью виртуальных экземпляров мы резервируем место для описания будущих находок и создаем структуру данных, которую впоследствии можно будет дополнить новыми информационными модулями, связанными с организмами. В качестве даты находок виртуальных экземпляров (компонент *когда*) можно использовать дату опубликования списка.



Рис. 12. Структуры баз данных для хранения списка таксонов: а) очевидная, но не оптимальная структура

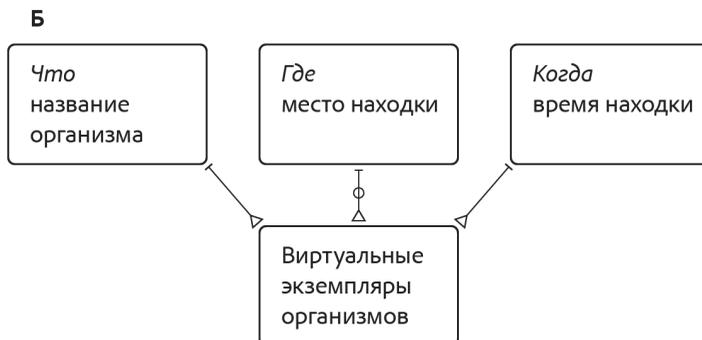


Рис. 12. Структуры баз данных для хранения списка таксонов: б) оптимальная структура с виртуальными экземплярами организмов, которые связаны с остальными компонентами

В некоторых таксономических списках приведены литературные источники, откуда извлечена информация о находке. Это еще один компонент (*откуда*), который также необходимо связать с виртуальными экземплярами. Тогда датой обнаружения организма является дата выхода соответствующей публикации.

Таким образом, любой список таксонов можно описать таблицей с виртуальными экземплярами, к которой присоединены таблицы с элементарными компонентами данных. Такое описание может быть дополнено иными информационными модулями, содержащими данные об организмах.

**7.2. Морфологическая и молекулярно-генетическая коллекция организмов, собранных во время морских экспедиций.** Как и в предыдущем примере, начинать проектирование структуры базы данных следует с таблицы особей (рис. 2). В данном случае каждая пойманная особь помещена в коллекцию. Следовательно, таблицу особей необходимо связать с таблицей с коллекционными экземплярами, а ее с таблицей с информацией о коллекциях. Для описания коллекционных экземпляров используем стандартный минимальный перечень полей (рис. 11).

Организмы пойманы во время морских экспедиций в конкретных точках. Такие точки в научном флоте называются станциями: у них есть географические координаты и они относятся к определенной акватории. Значит, таблицу особей следует связать с таблицей с данными о станциях, а ее с таблицей с названиями географических мест (рис. 8). В зависимости от способа лова организмы на станциях могут быть собраны с определенной глубины (батометром) или со

слоя глубин (сетью). Эти глубины хранятся в соответствующих полях (*height* на рис. 8). Таким образом, первичная структура базы данных включает таблицу особей и четыре компонента: *что, где, когда и куда* (рис. 13).



Рис. 13. Структура базы для хранения данных об организмах, собранных во время научных рейсов и депонированных в биологическую коллекцию

Морская экспедиция называется рейсом. Рейс имеет номер и связан с научно-исследовательским судном (НИС), на котором проходит экспедиция. Например, 86 рейс НИС «Профессор Водяницкий». Значит, совокупность станций, выполненных в рейсе, необходимо связать с номером рейса, а его – с судном. Учтем, что у каждой станции есть название, чаще всего – порядковый номер. Его следует сохранить в таблице с описанием географической точки в дополнительном поле. В другом поле будет записан идентификатор рейса; в третьем – дана ссылка на методику сбора проб. Таким образом, таблица географических точек (станций) состыкована с модулем данных о рейсах.

Наконец, морфометрические данные об особях (размеры, объемы тела) и молекулярно-генетические последовательности также хранятся в отдельной таблице. Но записи этой таблицы связаны не с особями, а с коллекционными экземплярами. Потому что коллекционные экземпляры – это организмы, специально подготовленные для длительного хранения, которым сопоставлен уникальный коллекционный

номер. Это позволяет использовать их повторно для уточнения и получения новых данных. Тогда как организм, не помещенный в коллекцию, чаще всего теряется, а значит, теряется и вся потенциальная информация о нем. Таким образом, таблица с коллекционными экземплярами используется для присоединения новых информационных модулей, в данном случае – таблицы с морфометрическими и молекулярно-генетическими данными.

Как можно увидеть, использование универсальных компонент помогло последовательно спроектировать базу данных (рис. 13).

**8. Заключение.** Минимальная структура базы для хранения данных о разнообразии организмов строится из шести универсальных компонент. Каждая из них описывает один из аспектов информации об организме: названии, месте и времени находки; исследователях, которые имеют отношение к обнаружению и анализу организма; литературных источниках, откуда были взяты или где были опубликованы факты об организме; биологической коллекции, куда помещен организм.

Ядром базы является таблица с характеристикой экземпляров организмов (особей), которая соединяется с таблицами компонент. Атрибуты связей между особью и компонентами описывают промежуточные таблицы.

В базе данных могут присутствовать не все компоненты. Состав компонент определяется задачами, для решения которых предназначена база. Минимально возможная структура базы данных состоит всего из двух таблиц: с характеристикой особи и с деревом таксономических названий.

Если стандартных компонент не хватает для описания совокупности данных, то используют дополнительные таблицы. Они присоединяются к промежуточным таблицам и не нарушают фундаментальную структуру базы.

Выделенные компоненты позволяют биологам и программистам использовать общий язык для описания объектов предметной области. Компоненты стандартизируют структуру баз данных. Их применение решает проблему совместимости данных из разных баз и помогает генерировать очевидные метаданные с описанием содержимого базы.

### Литература

1. *Байков К.С. и др.* Электронные коллекции и проблемы биоразнообразия // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Вторая Всероссийская науч. конф. 2000. Т. 26. С. 58–65.
2. *Шитиков В.К., Розенберг Г.С.* Оценка биоразнообразия: попытка формального обобщения // Количественные методы экологии и гидробиологии 2005. С. 91–129.

3. *Скарлато О.А., Старобогатов Я.И., Лобанов А.Л., Смирнов И.С.* Биоразнообразии и возможности его анализа с применением компьютерных банков данных // Биоразнообразие: Степень таксономической изученности. 1994. С. 20–43.
4. *Запевалин А.А.* Компьютерные структуры данных для решения проблем биоразнообразия // Экология моря. 2001. Вып. 57. С. 103–108.
5. *Seregin A.P.* The largest digital herbarium in Russia is now available online! // *Taxon*. 2018. vol. 67. no. 2. pp. 465–467.
6. *Pugachev O. et al.* Creation of information retrieval system on the unique research collections of the Zoological Institute RAS // *Information Technologies in the Research of Biodiversity*. Springer Proceedings in Earth and Environmental Sciences. 2019. pp. 57–65.
7. *Костина Н.В.* Информационная система REGION: 25 лет развития и практического применения // *Известия Самарского научного центра Российской академии наук*. 2015. Т. 17. № 4-1. С. 15–24.
8. *Лашин С.А. и др.* Информационная система по биоресурсным коллекциям институтов ФАНО России // *Вавиловский журнал генетики и селекции*. 2018. № 22(3). С. 386–393.
9. *Шапков М.П., Чадин И.Ф., Иванова Н.В.* Методические рекомендации по стандартизации данных для публикации через глобальный портал gbif.org и подготовке статьи о данных // *Труды Кольского научного центра РАН*. 2017. № 6-5(8). С. 22–35.
10. *Wieczorek J. et al.* Darwin core: an evolving community-developed biodiversity data standard // *PLoS One*. 2012. vol. 7. e29715.
11. *Jones K.E. et al.* PanTHERIA: a species-level database of life history, ecology, and geography of extant and recently extinct mammals // *Ecology*. 2009. vol. 90. no. 9. pp. 2648.
12. *Дмитриева Е.В. др.* Электронная коллекция паразитов рыб Мирового океана Института морских биологических исследований им. А.О. Ковалевского // *Морской биологический журнал*. 2016. Т. 1. № 3. С. 27–31.
13. *Гребенников К.А.* Изучение биоразнообразия заповедников России в цифровой эпохе: опыт и перспективы // *Nature Conservation Research*. Заповедная наука. 2016. Т. 1. № 2. С. 1–10.
14. *Лях А.М.* Структура базы данных для виртуальных биологических коллекций // *Системы контроля окружающей среды*. 2019. № 3 (37). С. 29–34
15. *Schneider F.D. et al.* Towards an ecological trait-data standard // *Methods in Ecology and Evolution*. 2019. vol. 10. pp. 2006–2019.
16. *Specht A. et al.* A story of data won, data lost and data re-found: the realities of ecological data preservation // *Biodiversity Data Journal*. 2018. vol. 6. e28073.
17. *Blair J. et al.* Towards a catalogue of biodiversity databases: An ontological case study // *Biodiversity Data Journal*. 2020. vol. 8. e32765.
18. *Parr C.S., Guralnik R., Cellinese N., Page R.D.M.* Evolutionary informatics: unifying knowledge about the diversity of life // *Trends in Ecology and Evolution*. 2012. vol. 27. no. 2. pp. 94–103.
19. *La J.S., Williams K.J., Moritz C.* Biodiversity analysis in the digital era // *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*. 2016. vol. 371. no. 1702. pp. 2150337.
20. *Лях А.М., Лелеков С.Г.* Модульная структура таксономической базы данных // *Электронные информационные системы*. 2018. № 2. С. 59–70.
21. *Лобанов А.Л. и др.* Эволюция стандарта ZOOCOD – концепция отражения зоологических иерархических классификаций в плоских таблицах реляционных баз данных // *Труды 10-й Всероссийской научной конференции «Электронные биб-*

- лиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2008. С. 326–332.
22. *Jackson J.B.C., Johnson K.G.* Measuring past biodiversity // *Science*. 2001. 293. pp. 2401–2404.
  23. *Каменский П.А., Сазонов А.Э., Федянин А.А., Садовничий В.А.* Биологические коллекции: стремление к идеалу // *Acta Naturae*. 2016. Т. 8. № 2(29). С. 6–10.
  24. *Чеченкина Т.В., Сотникова М.В.* Можно ли оценить бесценное? К вопросу о сравнении стоимости научных коллекций // *Управление наукой и наукометрия*. 2019. Т. 14. № 4. 17 p.
  25. *Sabaj M.H.* Standard symbolic codes for institutional resource collections in herpetology and ichthyology: an Online Reference. Version 6.5. URL: [asih.org/sites/default/files/documents/symbolic\\_codes\\_for\\_collections\\_v6.5\\_2016.pdf](http://asih.org/sites/default/files/documents/symbolic_codes_for_collections_v6.5_2016.pdf) (дата обращения: 22.06.2020).
  26. *Лях А.М.* Уникальные семантические номера экземпляров биологических коллекций // *Системы контроля окружающей среды*. 2019. № 2(36). С. 43–50.
  27. *Guntsch A. et al.* Actionable, long-term stable and semantic web compatible identifiers for access to biological collection objects // *Database*. 2017. vol. 2017. 9 p.
  28. *Касаткин М.В., Иванов А.П.* Коллекция костей стеллеровой коровы *Hydrodamalis gigas* в Государственном биологическом музее имени К.А. Тимирязева // *Зоологические исследования*. 2018. № 20. С. 60–66.
  29. *Michener W.K.* Meta-information concepts for ecological data management // *Ecological Informatics*. 2006. vol. 1. no. 1. pp. 3–7.
  30. *Berendsohn W.G., Seltman P.* Using geographical and taxonomic metadata to set priorities in specimen digitization // *Biodiversity Informatics*. 2010. vol. 7. no. 2. pp. 120–129.

**Лях Антон Михайлович** — канд. биол. наук, старший научный сотрудник, отдел экологической паразитологии, Федеральное государственное бюджетное учреждение науки Федеральный исследовательский центр «Институт биологии южных морей имени А.О. Ковалевского РАН» (ФИЦ ИнБИОМ). Область научных интересов: таксономические информационные системы и базы данных, компьютерные определители организмов, морфологический анализ биологических форм. Число научных публикаций — 50. [me@antonlyakh.ru](mailto:me@antonlyakh.ru); пр. Нахимова, 2, 299011, Севастополь, Россия; р.т.: +7(8692)55-09-12.

**Поддержка исследований.** Работа выполнена по теме государственного задания ФИЦ ИнБИОМ № АААА-А18-118020890074-2 и при поддержке РФФИ и города Севастополя по проектам № 18-44-920024 и № 20-44-920004.

A. LYAKH  
**MINIMAL STRUCTURE OF THE DATABASE FOR STORING  
ORGANISMS' BIODIVERSITY DATA**

*Lyakh A. Minimal Structure of the Database for Storing Organisms' Biodiversity Data.*

**Abstract.** To date, a huge amount of data on organisms diversity has been accumulated. Databases help to store and use these data for scientific purposes. There exists several dozens of databases for storing biodiversity data that were described in publications. Each has an original structure which badly correlates with the structures of other databases. This complicates data exchange and the formation of big biodiversity data array.

The cause of this situation is the lack of the formal definitions of universal data components, which allow to build the database with any data on the diversity of organisms. The analysis of publications and author's experience show that such universal components are present in the characteristics of any organisms. For example, it is an organism taxonomic name and a location where it was found. There are six such components and they answer to one of the six questions: *what*, where, when, who, where from and where to. *What* determines the name of an organism; *where* determines the location where it was found; *when* indicates the date of finding; *who* enumerates the persons, who found and analyzed an organism; *where* from refers to publications, where data about an organism are extracted or published; *where to* shows in which biological collection an organism is put in.

Each component corresponds to a separate database table. These tables are linked to the table with data about organism (individual) and they are not linked with each other. Attributes of the links between the organism table and the component tables are stored in intermediate tables. They are used, for example, to store bibliographic facts, descriptions of collection items or geographical points. They also act as docking stations to which tables with any other information are attached.

The creation of any database about the diversity of living organisms begins with the definition of the table of organism specimens. It must be used even if there is no explicit data on organisms. In that case virtual organisms should be introduced and the other components should be linked with them by means of intermediate tables. The latter are docked to other data. Minimal structures of all the tables, links between them and examples of databases construction are described in the work.

**Keywords:** Data Components, Organism Data, Geographical Point, Taxonomic Name, Bibliographic Record, Biological Collection, Collection Item, Metadata.

**Lyakh Anton** — Ph.D., Senior Researcher, Department of Ecological Parasitology, A.O. Kovalevsky Institute of Biology of the Southern Seas of RAS (IBSS). Research interests: taxonomical informational systems and databases, computer identification keys, morphological analysis of biological shapes. The number of publications — 50. me@antonlyakh.ru; 2, Nahimova av., 299011, Sevastopol, Russia; office phone: +7(8692)55-09-12.

**Acknowledgements.** This work was supported by the scientific research theme of the Institute of Biology of the Southern Seas No AAAA-A18-118020890074-2 and by RFBR and Government of the Sevastopol according to the research project No 18-44-920024 and 20-44-920004.

## References

1. Baikov K.S. et al. [Electronic collections and biodiversity issues] *Elektronnyye biblioteki: perspektivnyye metody i tehnologii, elektornnyye kollektsii*. Vtoraya Vseross. nauch. konf. [Electronic libraries: advanced methods and technologies, electronic col-

- lections. Second all-Russian scientific conference]. 2000. Issue 26. pp. 58–65. (In Russ.).
2. Shitikov V.K., Rozenberg G.S. *Otsenka bioraznoobraziya: popytka formalnogo obobshcheniya*. [Biodiversity assessment: an attempt of formal generalization]. *Kolichestvennye metody ekologii i gidrobiologii* [Quantitative methods of ecology and hydrobiology] 2005. pp. 91–129. (In Russ.).
  3. Skarlato O.A., Starobogatov Ya.I., Lobanov A.L., Smironov I.S. *Bioraznoobrazie I vozmozhnosti ego analiza s primeneniem komputernykh bankov dannykh*. [Biodiversity and the possibilities for its analysis using computer data banks]. *Bioraznoobrazie: Stepen taxonomicheskoy izuchennosti*. 1994. pp. 20–43. (In Russ.).
  4. Zapevalin A.A. [Computer Data Structures for Biodiversity Problems Solving]. *Ecologiya Morya – Ecology of the Sea*. 2001. vol. 57. pp. 103–108. (In Russ.).
  5. Seregin A.P. The largest digital herbarium in Russia is now available online! *Taxon*. 2018. vol. 67. no. 2. pp. 465–467.
  6. Pugachev O. et al. Creation of information retrieval system on the unique research collections of the Zoological Institute RAS. *Information Technologies in the Research of Biodiversity*. Springer Proceedings in Earth and Environmental Sciences. 2019. pp. 57–65.
  7. Kostina N.V. [Information system REGION: 25years of development and practical application]. *Izvestiya Samarskogo Nauchnogo centra RAN – News of Samara Scientific Center*. 2015. Issue 17. vol. 4-1. pp. 15–24. (In Russ.).
  8. Lashin S.A. et al. [An integrated information system on bioresource collections of the FASO of Russia]. *Vavilovskii Zhurnal Genetiki i Seleksii – Vavilov Journal of Genetics and Select*. 2018. vol. 22(3). pp. 386393. (In Russ.).
  9. Shashkov M.P., Chadin I.F., Ivanova N.V. [Guide to best practices on a data standardization for publication via the global portal gbif.org and preparing data paper]. *Trudy Kolskogo Nauchnogo Centra RAN – Proc. of the Kolsky Scientific Center of RAS*. 2017. vol. 6-5(8). pp. 22–35. (In Russ.).
  10. Wicczorek J. et al. Darwin core: an evolving community-developed biodiversity data standard. *PLoS One*. 2012. vol. 7. e29715.
  11. Jones K.E. et al. PanTHERIA: a species-level database of life history, ecology, and geography of extant and recently extinct mammals. *Ecology*. 2009. vol. 90. no. 9. pp. 2648.
  12. Dmitrieva Ye.V. et al. [The electronic collection of the World Ocean fish parasites maintained by Kovalevsky Institute of Marine Biological Research RAS]. *Morskoy biologicheskij zhurnal – Marine Biological Journal*. 2016. Issue 1. vol. 3. pp. 27–31. (In Russ.).
  13. Grebennikov K.A. [Study of biodiversity of nature reserves of the Russia in the digital age: experience and perspectives]. *Nature Conservation Research. Zapovednaya nauka – Nature Conservation Research. Conservation science*. 2016. Issue 1. vol. 2. pp. 1–10. (In Russ.).
  14. Lyakh A.M. [Database structure for virtual biological collections]. *Sistemy Controlya Okruzhayushey Sredy – Environmental Control Systems*. 2019. vol. 3(37). pp. 29–34. (In Russ.)
  15. Schneider F.D. et al. Towards an ecological trait-data standard. *Methods in Ecology and Evolution*. 2019. vol. 10. pp. 2006–2019.
  16. Specht A. et al. A story of data won, data lost and data re-found: the realities of ecological data preservation. *Biodiversity Data Journal*. 2018. vol. 6. e28073.
  17. Blair J. et al. Towards a catalogue of biodiversity databases: An ontological case study. *Biodiversity Data Journal*. 2020. vol. 8. e32765.

18. Parr C.S., Guralnik R., Cellinese N., Page R.D.M. Evolutionary informatics: unifying knowledge about the diversity of life. *Trends in Ecology and Evolution*. 2012. vol. 27. no. 2. pp. 94–103.
19. La J.S., Williams K.J., Moritz C. Biodiversity analysis in the digital era. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*. 2016. vol. 371. no. 1702. pp. 2150337.
20. Lyakh A.M., Lelekov S.G. [The modular structure of a taxonomic database]. *Elektronnyye Informatsionnyye Sistemy – Electronic Information Systems*. 2018. vol. 2. pp. 59–70. (In Russ.).
21. Lobanov A.L. et al. [ZOOCOD standard evolution – the concept of the reflection of zoological hierarchical classifications in flat tables of relational databases]. *Trudy 10 Vserossiyskoy nauchnoy konferentsii “Elektronnyye biblioteki: perspektivnyye metody i tekhnologii, elektronnyye kolleksii” – RCDL’2008*. [Proc. of 10<sup>th</sup> all-russian scientific conference “Electronic libraries, innovative methods and technologies, electronic collections”]. 2008. pp. 326–332. (In Russ.).
22. Jackson J.B.C., Johnson K.G. Measuring past biodiversity. *Science*. 2001. vol. 293. pp. 2401–2404.
23. Kamenski P.A., Sazonov A.E., Fedyanin A.A., Sadovnichy V.A. [Biological collections: chasing the ideal]. *Acta Naturae*. 2016. Issue 8. vol. 2(29). pp. 6–10. (In Russ.).
24. Chechenkina T.V., Sotnikova M.V. [Is it possible to estimate the value of invaluable? More on the ranking of scientific collections by value]. *Upravlenie naukoj i naukometriya – Science Governance and Scientometrics*. 2019. Issue 14. vol. 4. (In Russ.).
25. Sabaj M.H. Standard symbolic codes for institutional resource collections in herpetology and ichthyology: an Online Reference. Version 6.5. Available at: [asih.org/sites/default/files/documents/symbolic\\_codes\\_for\\_collections\\_v6.5\\_2016.pdf](http://asih.org/sites/default/files/documents/symbolic_codes_for_collections_v6.5_2016.pdf) (accessed: 22.06.2020).
26. Lyakh A.M. [Unique semantic numbers of biological collection items for databases]. *Sistemy Controlya Okruzhayushey Sredy – Environmental Control Systems*. 2019. vol. 2(36). pp. 43–50. (in Russ.).
27. Guntsch A. et al. Actionable, long-term stable and semantic web compatible identifiers for access to biological collection objects. *Database*. 2017. vol. 2017. 9 p.
28. Kasatkin M.V., Ivanov A.P. Collection of bones of Steller’s sea cow *Hydrodamalis gigas* in the Tmiryazev State Biological Museum (Moscow). *Zoologicheskie Issledovaniya – Zoological Studies*. 2018. vol. 20. pp. 60–66. (In Russ.).
29. Michener W.K. Meta-information concepts for ecological data management. *Ecological Informatics*. 2006. vol. 1. no. 1. pp. 3–7.
30. Berendsohn W.G., Seltman P. Using geographical and taxonomic metadata to set priorities in specimen digitization. *Biodiversity Informatics*. 2010. vol. 7. no. 2. pp. 120–129.